Project Part 3: Make Predictions
CSE 446: Machine Learning

University of Washington

Deadline: December 5, 2017
(late submissions will not be accepted under any circumstances)

November 7, 2017: updates in red. November 18, 2017: updates in blue.

For each part of the project, your team will be evaluated as a whole; you will share a grade. The third task for your team to complete is to apply machine learning algorithms to make predictions on the competition datasets.

There are seven competition datasets (listed in no particular order):

name	creator team's ID	N	notes
miri_playlist	8	10,503	regression: $d = 8$
washington_flights	25	10,044	regression: $d = 9$
breastcancer	42	800	classification: $d = 12$
mushroom edibility	7	*5,384	classification: $d = 143$
FraudBGone	22	2,993	classification: text
VANGOGHORNO	5	481	classification: images
lan-gauge	13	640	classification: audio

*Including a pre-defined development dataset.

These are available as a tarball on Canvas.

(Congratulations to the teams whose datasets were selected! In addition to bragging rights, your prize is that you might need to provide assistance to classmates with questions about the data. With great power comes great responsibility: you must not, under any circumstances, share the test data with anyone except for course staff.)

Your team is required to make predictions on **twice as many datasets as there are team members**. That is, three-member teams will make predictions on six datasets; two-member teams will make predictions on four datasets. If your dataset is one of the competition datasets, you may not choose it. Taken together, these rules imply that a three-member team whose dataset from P1 was chosen for the competition *has no choice but to make predictions on all six other datasets*. From here on, we will refer to your team's "chosen datasets," regardless of whether you actually had a choice.

For each of your team's chosen datasets, you must learn a predictor (i.e., a classifier or a regressor) from the training data. You will submit *one* set of predictions on the test set for each of your chosen datasets. Since you get only one shot, you must take hyperparameter selection very seriously!

You do not need to use the same approach for each dataset (though you are *allowed* to use the same approach for more than one dataset). For each chosen dataset, your team must apply *one* approach to make predictions on its test set. (The test sets will be made available to you shortly before the final deadline.)

In addition to submitting predictions, you must compose a single pdf explaining your approach(es). If you explored a range of options before settling on a final approach, discuss your exploration. For each dataset, explain the final approach you used: (i) your features; (ii) your choice of learning algorithm/model/loss function; (iii) your optimization technique; (iv) what hyperparameters your approach required, and how you tuned them; (v) how you estimated performance.

For full credit, your document should leave absolutely no confusion about which approach corresponds to which dataset. You should provide enough detail that another student who paid attention in CSE 446 could replicate your approach and get very similar predictions. If you found something interesting (e.g., a strong effect from hyperparameter tuning or a feature that made a big difference), feel free to discuss and illustrate with tables and charts. Altogether, aim to be concise; your report should not be more than three pages (excluding tables and figures, which you may include in an appendix). If you do include tables and/or figures, make sure each one is clearly explained in its caption and referenced in the text; illustrations that do not make your report more clear will result in a loss of points.

The report must include an exhaustive list of existing libraries and software packages that you used in implementing your solution. It is perfectly fine to use existing tools, as long as you acknowledge them. Failure to acknowledge will result in a serious penalty; ask the course staff if you have any questions.

Note that the report constitutes 15% of your grade.

Turn in a single gzipped tarball containing the pdf writeup and your predictions on all test sets. Your output filenames should be `TEAM.DATASET.test-yhat`, one per competition test set, formatted with one prediction per line, parallel to the `test-x` file (i.e., the first line of your file is your prediction on the input in the filename on the first line of the `test-x` file, and so on). Here, `TEAM` is your team's ID and `DATASET` is the ID for the dataset.

Note that the predictions constitute 5% of your grade. This 5% will be scaled based on how well your group performs in the Kaggle competitions (see below). There will be **absolutely no partial credit** if you fail to submit properly formatted predictions. This includes filenames that fail to *exactly* follow the requirements above.

The tarball should also contain source code that will allow us to reproduce your results, and to check that your code does what your writeup says it does.

Kaggle

Your team was required to make predictions on **twice as many datasets as there are team members**. Now, you will **compete in a Kaggle competition for each of those datasets**.

First, please create a Kaggle account at `https://www.kaggle.com/`. The URL of the Kaggle competition for each dataset can be found in the table below:

name	competition URL
miri_playlist	https://www.kaggle.com/t/76472283c03147ac97874f5f3ac033ff
washington_flights	https://www.kaggle.com/t/81a557404cfa4ed7a8d23fa13ab17feb
breastcancer	https://www.kaggle.com/t/5b7000b7c00243418a16a4cdd8308622
mushroom edibility	https://www.kaggle.com/t/42ea23e545224d14b3f923c0d290cbfb
FraudBGone	https://www.kaggle.com/t/fa92423dafe14fd4964800489f92ba01
VANGOGHORNO	https://www.kaggle.com/t/88d840642dcd49c98c82c5630f1c8441
lan-gauge	https://www.kaggle.com/t/683c105a7391445fa2b48b417ba18a25

For each Kaggle competition you will participate in, create a group which is comprised of your project group members. You can do this under the "team" tab on the Kaggle competition website. You can name your group whatever you would like, but use the same group name across all competitions that you register for. Once you have created your group, report your Kaggle account information to us through this Google Form: `https://goo.gl/forms/ChpkwYbeZj68HGv72`.

In order to create Kaggle submissions, please use the `to_kaggle.py` script available on the Kaggle competition pages. To create a Kaggle submission, you can run: `python3 to_kaggle.py <b|r> <test-x> <test-yhat> <out-filepath>`. The first argument specifies whether the predictions are for a binary classification or regression problem. The second argument is the filepath of the dataset's `test-x` file. The third argument is the filepath of your `test-yhat` file for the dataset. The final argument is the output filepath where the Kaggle submission file will be created. For example, to create a Kaggle submission for miri_playlist, you can run: `python3 to_kaggle.py r test-x 30.miri_playlist.test-yhat sub.kaggle`. On the Kaggle competition page, click **submit predictions** and upload the Kaggle submission file generated by `to_kaggle.py`. This **only works with UTF-8 encoded files** and some of the `test/train-x/y` files that were submitted used different file encodings. To make it easier to use the `to_kaggle.py` script, we have uploaded datasets with UTF-8 encoded `test/train-x/y` files to Canvas.

Each day, your team will be allowed two submssions. For each of these submissions, a public score will be displayed on the leaderboard. Note that the public score is calculated using only a portion of the entire test set. The final leaderboard will be calculated using the other portion of test data. Consider carefully the implications of this ranking scheme! Before the end of the competitions, please mark which submission you would like to be used for evaluation (on the private test data).

You are not allowed to use test data in **any** way to help train your models. This means you are not allowed to use transfer learning or domain adaptation, or to try to label the test data for use as training data. We will be running your code to ensure that your Kaggle scores align with how well your models perform.

Again, note that the competitions constitute 5% of your grade. There will be **absolutely no partial credit** if you fail to submit predictions to Kaggle.